A short note on decoupling the efficiency and missing/causal assumption

Yen-Chi Chen
University of Washington
February 12, 2021

When we learn the efficiency theory from a conventional semi-parametric theory course, we may often think of the efficiency theory and the missing data assumption are highly relevant. In this note, we will argue that they are actually **decoupled**. You can construct efficiency theory (and double robustness) even if you do not make any missing data assumption. A similar result also applies to causal inference problem. In short,

> *the efficiency theory and double-robustness is a property of an estimator (observed-data quantity), it has nothing to do with the missing/causal assumption.*

1 A simple example

Consider a very simple example where we have three variables X,Y,A such that X is always observed and Y may be missing and $A \in \{0,1\}$ is the binary variable indicating the presence of Y, i.e., $A=1$ if Y is observed.

Consider $\theta_0 = \mathbb{E}(Y)$. In general, θ_0 is not identifiable because Y may be missing and all three variables may be dependent.

A conventional way to identify θ_0 is via assuming the missing at random (MAR) assumption, which is

$$(MAR) \qquad Y \perp A|X$$

in the current setup. Under the MAR assumption,

$$\theta_0 = \mathbb{E}(Y) = \mathbb{E}\left(\frac{YA}{\pi(X)}\right) = \mathbb{E}(m_1(X)),$$

where $\pi(x) = P(A=1|X=x)$ is the propensity score and $m_1(x) = \mathbb{E}(Y|A=1,X=x)$ is the complete-case outcome regression function.

The quantity $\mathbb{E}\left(\frac{YA}{\pi(X)}\right)$ leads to the inverse probability weighting (IPW) estimator and the quantity $\mathbb{E}(m_1(X))$ leads to the regression adjustment (a.k.a. g-computation) estimator. Let

$$L_{\mathsf{IPW}}(X,Y,A) = \frac{YA}{\pi(X)}, \quad L_{\mathsf{RA}}(X,Y,A) = m_1(X)$$

be the corresponding linear forms of the IPW and regression adjustment (RA) estimators.

Moreover, we further have the following equality:

$$\mathbb{E}\left(\frac{YA}{\pi(X)}\right) = \mathbb{E}(m_1(X)) = \mathbb{E}\left(m_1(X) + \frac{A}{\pi(X)}(Y - m_1(X))\right) \tag{1}$$

and the last quantity leads to the linear form of the doubly robust (DR) estimator:

$$L_{\mathsf{DR}}(X,Y,A) = m_1(X) + \frac{A}{\pi(X)}(Y - m_1(X)).$$

In the conventional semi-parametric theory course, we have learned that $L_{\mathsf{DR}}(X,Y,A)$ leads to the efficient estimator and has the so-called doubly-robust property, i.e., we just need one of $\pi(x)$ or $m_1(x)$ to be correct to obtain a consistent estimator of θ_1.

Now we think about the following question: *what if MAR is incorrect but we still use the above estimators, what will happen?*

First, you can easily see that all three linear forms have the same expectation:

$$\begin{aligned}\theta_1 &= \int yp(y|A=1,x)dyp(x)dx\\ &= \mathbb{E}(L_{\mathsf{IPW}}(X,Y,A))\\ &= \mathbb{E}(L_{\mathsf{RA}}(X,Y,A))\\ &= \mathbb{E}(L_{\mathsf{DR}}(X,Y,A)).\end{aligned}$$

And if MAR is not correct, $\theta_1 \neq \theta_0$ in general. So these estimators lead to a biased estimate of θ_0.

However, they are still consistent estimators of θ_1! Here is a more interesting fact: we can rewrite $L_{\mathsf{DR}}(X,Y,A)$ as

$$\begin{aligned}L_{\mathsf{DR}}(X,Y,A) &= m_1(X) + \underbrace{\frac{A}{\pi(X)}(Y - m_1(X))}_{\mathbb{E}(\cdot|X,A)=0}\\ &= \frac{AY}{\pi(X)} + \underbrace{m_1(X)\left(1 - \frac{A}{\pi(X)}\right)}_{\mathbb{E}(\cdot|X)=0}\end{aligned}$$

so $L_{\mathsf{DR}}(X,Y,A)$ still leads to a *doubly robust estimator of* θ_1! Also, you can show that it is the efficient estimator of θ_1. All these properties are retained even if the missing data assumption is incorrect.

In a sense, this is like the problem of model mis-specification problem of a parametric model. We are not inferring the true underlying distribution but a projection of the true distribution to certain space.

2 Remarks

- **Binary treatment effect.** In the binary treatment problem, you can see that the same reasoning also applies. The usual IPW/RA/DR estimators are estimating the quantity

$$\theta_{\mathsf{ATE},1} = \int (m_1(x) - m_0(x))p(x)dx,$$

which is and observed-data quantity and will equal the average treatment effect under the strong ignorability condition. All the efficiency theory is for this quantity ($\theta_{\mathsf{ATE},1}$) and has nothing to do with the ignorability condition.

In fact, this result is pretty straight forward if we think of the causal effect using the do operator and the directed acyclic graph (DAG); see http://faculty.washington.edu/yenchic/short_note/note_adjustment.pdf. The do operator approach directly define an observed quantity as a causal effect so all the nice property is about this observed quantity.

- **Universality for any missing/causal assumption.** This result applies to any missing and causal assumptions, not limited to the MAR/strong ignorability. The assumption on missingness of causality motivates us to derive an observed-data quantity as an estimator to the quantity/parameter of interest. But the efficiency of such estimator is irrelevant to the missing/causal assumption we made.

- **Viewing it as bias-variance decomposition.** We may view the problem as a bias-variance decomposition. Let θ_0 be the parameter of interest that may depend on unobserved quantity and θ_1 is the quantity that our estimator $\widehat{\theta}$ is converging to. Then we can describe the three quantities using the following diagram:

$$\theta_0 \qquad \underbrace{\longleftrightarrow}_{\text{missing/causal assumption}} \qquad \theta_1 \qquad \underbrace{\longleftrightarrow}_{\text{efficiency theory}} \qquad \widehat{\theta}. \tag{2}$$

  The missing/causal assumption determines the difference between θ_1 and θ_0, so it is related to the bias (assumption bias) and has nothing to do with the distance between $\widehat{\theta}$ and θ_1. The efficiency theory describes the closeness between θ_1 and $\widehat{\theta}$ and has nothing to do with the difference $\theta_0 - \theta_1$. The efficiency theory is relevant to the variance of the estimator but not the bias.

  Similar to model mis-specification problems, the assumption/model bias and the variance of estimating the model are often decoupled. Here, the same scenario occurs–the assumption of missingness and causality and the efficiency theory are decoupled.

- **The interplay between missing/causal assumption and efficiency theory.** As we have argue, missing/causal assumption has nothing to do with the efficiency theory. However, the assumption and the efficiency theory will be related in our data analysis pipeline. Here is the reason.

  When we have the data and set a parameter of interest, we need to make a decision on the missing/causal assumption when there is incompleteness of the data. Different choices in this stage (assumptions) will lead to different estimators. After we have decided the assumption and find an observed-data quantity as the estimator, we then study the efficiency of such estimator. So the two concepts, missing/causal assumptions and efficiency theory, are related in the data analysis pipeline.

  Here is a metaphor for their relation. Think about a factory that can produce various products. The efficiency theory is about how fast the factory can produce a particular product and the missing/causal assumption determines which product is popular on the market. The two concepts (efficiency of producing and the marketing strategy) are independent on their own. But if we are thinking about optimizing the total income, they have to be aligned with each other.